直線回帰とロジスティック回帰の違い

図 1 のように，従属変数が 2 値変数（2 種類の値のどちらかしかとらない）場合に，直線回帰を行うのは不適切である。例えば，0/1 データの場合，$[0, 1]$ の値をとるということは，1 の事象が起きる確率として解釈できるが，0 未満の値や 1 を超える値の解釈ができない。

ある事象が発生する（従属変数が 1 になる）確率を P としたとき，$\frac{P}{1-P}$ はオッズ*1，その対数をとった $\log\left(\frac{P}{1-P}\right)$ はロジットまたは対数オッズと呼ばれる。

(1) 式のようにロジットが独立変数の線形結合式 $\lambda = b_0 + b_1\ x_1 + b_2\ x_2 + \ldots + b_p\ x_p$ で表せるとするのがロジスティックモデルである。λ を線形予測子 linear predictor ともいう。

$$\log\left(\frac{P}{1-P}\right) = b_0 + b_1\ x_1 + b_2\ x_2 + \ldots + b_p\ x_p = \lambda \tag{1}$$

*1 オッズというのはギャンブルで「見込み」を表すために使われるものである。$\frac{\text{成功の回数}}{\text{失敗の回数}}$ で計算されるもので，日本の公営競馬，競輪などにおけるオッズとは定義が違う。分子と分母をそれぞれ「成功の回数と失敗の回数の合計」で割れば，$\frac{\text{成功の確率}}{\text{失敗の確率}} = \frac{\text{成功の確率}}{1-\text{成功の確率}}$ となる。成功とは「注目している事象が起きること」とする。

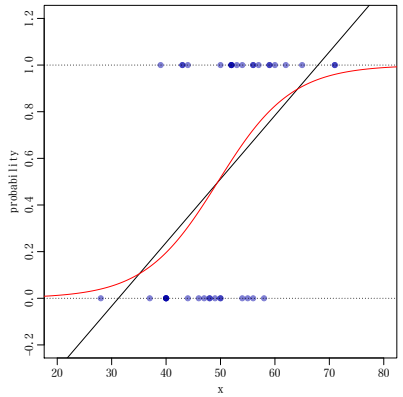

図 1　ロジスティック回帰分析と重回帰分析の違い

(1) 式の両辺の逆対数をとった (2) は，従属変数が 1 になる確率と 0 になる確率の比をとったもの（オッズ）である。

$$\frac{P}{1-P} = \exp(b_0 + b_1\ x_1 + b_2\ x_2 + \ldots + b_p\ x_p) \tag{2}$$

x_1 以外の独立変数が同じであるという条件下で，$x_1 = u + 1$ のときのオッズと $x_1 = u$ のときのオッズの比をとると (3) 式のようになる。これは，オッズ比と呼ばれ，x_i が 1 単位増えたときに，オッズは $\exp(b_i)$ 倍になることを意味する。

$$\frac{P_{x_1=u+1}/(1-P_{x_1=u+1})}{P_{x_1=u}/(1-P_{x_1=u})} = \frac{\exp(b_0 + b_1\ (u+1) + b_2\ x_2 + \ldots + b_p\ x_p)}{\exp(b_0 + b_1\ u + b_2\ x_2 + \ldots + b_p\ x_p)} = \exp(b_1) \tag{3}$$

(1) 式を P について解くと (4) 式のようになる。

$$P = \frac{1}{1 + \exp(-\lambda)} \tag{4}$$

P は $[0, 1]$ の値をとり，λ を横軸，P を縦軸として描かれる曲線はロジスティック曲線と呼ばれる*2。

*2 同じような「S 字状曲線」を表すプロビットモデル（プロビット曲線）があるが，数学的に簡単に取り扱えるロジスティックモデルのほうがよく使われる。

ロジスティックモデルの種類

多重ロジスティックモデル　従属変数は 2 値データ

多項ロジスティックモデル　従属変数は 3 つ以上のカテゴリーを持つ名義尺度データ

順序ロジスティックモデル　従属変数は 3 つ以上のカテゴリーを持つ順序尺度データ（カテゴリーに順序関係がある）

累積ロジスティックモデル

比例オッズモデル

多重ロジスティックモデル

多重ロジスティックモデルは，従属変数が二値データである場合に適用される。

分析例は表 1 のようなデータである。

表 1　従属変数が二値データ

	x_1	x_2	x_3	y
1	62.4	51.7	51.8	0
2	48.3	47.9	43.8	0
3	61.1	37.9	48.6	0
4	48.4	42.3	37.4	0
5	48.0	57.1	61.1	1
6	50.0	53.2	43.9	1
7	48.0	36.0	34.9	0
8	51.8	53.0	68.9	1
9	36.5	32.4	35.2	0
10	33.1	47.5	43.8	0
:	:	:	:	:
99	55.9	55.7	61.5	1
100	68.7	64.8	59.0	1

分析結果は表 2 に示す。

表 2　多重ロジスティックモデルによる解析結果

	偏回帰係数	標準誤差	z 値	P 値
x1	0	0.03194	0.916	0.360
x2	0	0.03880	3.582	< 0.001
x3	0	0.03785	2.310	0.021
定数項	−13	2.65285	−4.836	< 0.001

ロジット λ は，元のデータを偏回帰係数により線形変換したものである。

$$\lambda = -12.82897 + 0.02926\ x_1 + 0.13895\ x_2 + 0.08743\ x_3$$

事象の生起確率 P は，ロジットを用いて計算される（表 3）。

$$P = \frac{1}{1 + \exp(-\lambda)}$$

表 3　分析結果

	x_1	x_2	x_3	y	ロジット	生起確率	判別結果
1	62.4	51.7	51.8	0	0.70992	0.67038	1
2	48.3	47.9	43.8	0	-0.93017	0.28289	0
3	61.1	37.9	48.6	0	-1.52546	0.17866	0
4	48.4	42.3	37.4	0	-2.26494	0.09407	0
5	48.0	57.1	61.1	1	1.85195	0.86436	1
6	50.0	53.2	43.9	1	-0.13522	0.46625	0
7	48.0	36.0	34.9	0	-3.37063	0.03323	0
8	51.8	53.0	68.9	1	2.07540	0.88849	1
9	36.5	32.4	35.2	0	-4.18118	0.01505	0
10	33.1	47.5	43.8	0	-1.43058	0.19301	0
:	:	:	:	:	:	:	:
99	55.9	55.7	61.5	1	1.92358	0.87254	1
100	68.7	64.8	59.0	1	3.34408	0.96591	1

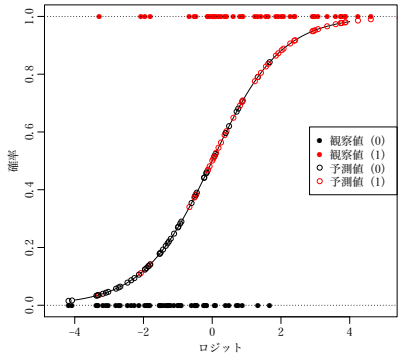

図 2　観察値とロジットと予測値（確率）

予測値（生起確率）が 0.5 以上の場合，事象が発生したとすれば，予測結果は表 4 のようにまとめられる。

正判別率は 81% である。

表 4　多重ロジスティックモデルによる判別結果

	0	1	合計
0	43	9	52
1	10	38	48

なお，線形判別分析によれば表 5 のようになる。

線形判別分析による正判別率は 80% である。

表 5　線形判別分析による判別結果

	0	1	合計
0	45	7	52
1	13	35	48

多項ロジスティックモデル

多項ロジスティックモデルは，従属変数が 3 つ以上のカテゴリーを持つ名義尺度データに適用できる。従属変数は多項分布に従うと仮定される。

表 6 に示したようなデータを使用する。

表 6　従属変数が名義尺度データ

	x_1	x_2	x_3	y
1	62.4	51.7	51.8	3
2	48.3	47.9	43.8	1
3	61.1	37.9	48.6	1
4	48.4	42.3	37.4	2
5	48.0	57.1	61.1	3
6	50.0	53.2	43.9	2
7	48.0	36.0	34.9	1
8	51.8	53.0	68.9	3
9	36.5	32.4	35.2	1
10	33.1	47.5	43.8	1
:	:	:	:	:
99	55.9	55.7	61.5	2
100	68.7	64.8	59.0	2

分析結果は表 7 のように，切片と偏回帰係数の推定値が 2 組（従属変数のカテゴリー数より 1 少ない）得られる。

表 7　多項ロジスティックモデルの結果

	偏回帰係数	標準誤差	z 値
(Intercept):1	10.15774	2.328	4.363
(Intercept):2	6.39061	2.007	3.184
x1:1	−0.05806	0.036	−1.628
x1:2	−0.01886	0.032	−0.594
x2:1	−0.06200	0.038	−1.616
x2:2	−0.00890	0.034	−0.263
x3:1	−0.08079	0.041	−1.965
x3:2	−0.08902	0.036	−2.478

切片と傾きから計算されるロジットは，判別される群の数より 1 つ少ない個数だけ計算される（表 8）。ロジット 1 は $y = 3$ のものに対する $y = 1$ のロジット，ロジット 2 は $y = 3$ のものに対する $y = 2$ のロジットを意味する。

表 8　多項ロジスティックモデルの結果

	x_1	x_2	x_3	y	$logit_1$	$logit_2$	p_1	p_2	p_3	判別結果
1	62.4	51.7	51.8	3	-0.85600	0.14297	0.16477	0.44742	0.38781	2
2	48.3	47.9	43.8	1	0.84462	1.15479	0.35799	0.48818	0.15384	2
3	61.1	37.9	48.6	1	0.33369	0.57509	0.33453	0.42586	0.23961	2
4	48.4	42.3	37.4	2	1.70313	1.77242	0.44368	0.47552	0.08080	2
5	48.0	57.1	61.1	3	-1.10616	-0.46136	0.16868	0.32144	0.50988	3
6	50.0	53.2	43.9	2	0.40921	1.06669	0.27823	0.53697	0.18480	2
7	48.0	36.0	34.9	1	2.31897	2.05855	0.53502	0.41235	0.05263	1
8	51.8	53.0	68.9	3	-1.70277	-1.19088	0.12259	0.20453	0.67289	3
9	36.5	32.4	35.2	1	3.18563	2.28073	0.69159	0.27981	0.02860	1
10	33.1	47.5	43.8	1	1.75192	1.44499	0.52379	0.38536	0.09085	1
:	:	:	:	:	:	:	:	:	:	:
99	55.9	55.7	61.5	2	-1.51034	-0.63349	0.12608	0.30301	0.57092	3
100	68.7	64.8	59.0	2	-2.61576	-0.73328	0.04706	0.30920	0.64373	3

3 つのカテゴリーの生起確率は，

$$\text{従属変数が}\begin{cases} 1 \text{ である確率 } p_1 = \dfrac{\exp(logit_1)}{1+\exp(logit_1)+\exp(logit_2)} \\ 2 \text{ である確率 } p_2 = \dfrac{\exp(logit_2)}{1+\exp(logit_1)+\exp(logit_2)} \\ 3 \text{ である確率 } p_3 = \dfrac{1}{1+\exp(logit_1)+\exp(logit_2)} \end{cases} \tag{5}$$

により計算される。

従属変数を予測するには，計算された確率が最も高いものであると判別すればよい。判別結果は表 9 に示す。

正判別率は 54% である。

表 9　多項ロジスティック判別による判別結果

	1	2	3	合計
1	14	15	2	31
2	8	21	10	39
3	4	7	19	30

なお，正準判別分析による判別結果は表 10 のようになる。

正準判別分析による正判別率は 54% である。

表 10　正準判別分析による判別結果

	1	2	3	合計
1	14	15	2	31
2	8	21	10	39
3	3	8	19	30

順序ロジスティックモデル

従属変数が 3 つ以上のカテゴリーを持ち，しかもそのカテゴリーに順序関係がある場合（順序尺度データ）に使用されるモデルである。

表 6 のデータを用いる。ただし，従属変数には $1 < 2 < 3$ のように順序がある，順序尺度だとする。

累積ロジスティックモデル（cumulative logistic model）では，まず，“1” を「反応あり」，“2” と “3” を「反応なし」として

$$\lambda_1 = \log\frac{\pi_1}{1-\pi_1} = \beta_{10} + \beta_{11}\ x_1 + \cdots + \beta_{1p}\ x_p + \varepsilon_1 \tag{6}$$

を当てはめる。

次に “1” と “2” を「反応あり」，“3” を「反応なし」として

$$\lambda_2 = \log\frac{\pi_2}{1-\pi_2} = \beta_{20} + \beta_{21}\ x_1 + \cdots + \beta_{2p}\ x_p + \varepsilon_2 \tag{7}$$

を当てはめる。

ここで，(6) 式と (7) 式の偏回帰係数において，$\beta_{10} \neq \beta_{20}$ であるがそのほかの偏回帰係数は $\beta_{1i} = \beta_{2i}$ であると考える場合と，偏回帰係数は全て異なる（$\beta_{1i} \neq \beta_{2i}$）と考える場合とがある。

前者は，2 つのモデルのオッズの間には比例関係がありその比例定数は $\exp(\beta_{10} - \beta_{20})$ となるので，このような累積ロジスティックモデルは特に比例オッズモデル（proportional odds model, POM）と呼ぶ。2 つのモデルのオッズの間には比例関係があり（ロジットの差 $\beta_{10} - \beta_{20}$ は一定），その比例定数は定数項の差を指数変換した値になる。

後者では，比例関係にはない。

図 3 のロジスティック曲線において，$(logit_1, p_1)$ と $(logit_2, p_2)$ がそれぞれ (6) 式と (7) 式によるものとする。ロジットに対応する生起確率に基づいて，そのデータが従属変数のどのカテゴリーに属するかは，以下のように予測される。

まず，赤の $logit_1$ に対する p_1 は “1” を「反応あり」としたものであるから，確率 p_1 は “1” である確率である。

次に，黒の $logit_2$ に対する p_2 は，“1” と “2” を「反応あり」としたものであるから，確率 p_2 は “1” または “2” である確率なので，$1 - p_2$ が “3” の確率である。

よって，$1 - \{(1 - p_2) + p_1\} = p_2 - p_1$ が “2” である確率である。

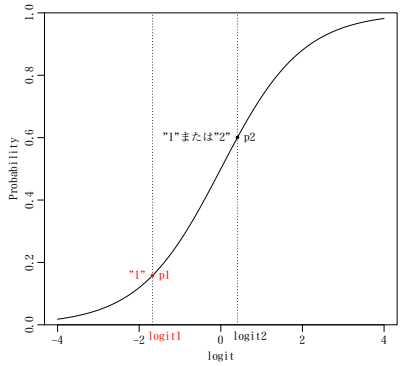

図 3　それぞれの確率

比例オッズモデル

表 11　比例オッズモデルの結果

	偏回帰係数	標準誤差	z 値
(Intercept):1	5.82201	1.369	4.252
(Intercept):2	7.90346	1.482	5.333
x1	−0.03717	0.023	−1.613
x2	−0.04146	0.025	−1.673
x3	−0.05854	0.026	−2.210

表 12 において，$logit_1$ と $logit_2$ の差は一定である。

表 12　比例オッズモデルの結果

	x_1	x_2	x_3	y	$logit_1$	$logit_2$	p_1	p_2	p_3	判別結果
1	62.4	51.7	51.8	3	-1.67314	0.40831	0.15801	0.44268	0.39932	2
2	48.3	47.9	43.8	1	-0.52319	1.55826	0.37211	0.45400	0.17390	2
3	61.1	37.9	48.6	1	-0.86536	1.21609	0.29622	0.47515	0.22863	2
4	48.4	42.3	37.4	2	0.07991	2.16136	0.51997	0.37676	0.10327	1
5	48.0	57.1	61.1	3	-1.90619	0.17526	0.12941	0.41429	0.45630	3
6	50.0	53.2	43.9	2	-0.81196	1.26949	0.30747	0.47318	0.21935	2
7	48.0	36.0	34.9	1	0.50232	2.58377	0.62300	0.30681	0.07019	1
8	51.8	53.0	68.9	3	-2.33406	-0.25261	0.08834	0.34884	0.56282	3
9	36.5	32.4	35.2	1	1.06146	3.14291	0.74297	0.21566	0.04137	1
10	33.1	47.5	43.8	1	0.05838	2.13983	0.51459	0.38012	0.10529	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
99	55.9	55.7	61.5	2	-2.16520	-0.08375	0.10292	0.37616	0.52093	3
100	68.7	64.8	59.0	2	-2.87190	-0.79045	0.05356	0.25851	0.68793	3

正判別率は 52% である。

表 13　比例オッズモデル判別による判別結果

	1	2	3	合計
1	14	15	2	31
2	9	22	8	39
3	2	12	16	30

累積ロジスティックモデル

表 14　累積ロジスティックモデルの結果

	偏回帰係数	標準誤差	z 値
(Intercept):1	5.47001	1.674	3.267
(Intercept):2	8.39935	1.858	4.521
x1:1	−0.04906	0.028	−1.757
x1:2	−0.03050	0.029	−1.054
x2:1	−0.06385	0.031	−2.088
x2:2	−0.02197	0.031	−0.709
x3:1	−0.01619	0.032	−0.503
x3:2	−0.09341	0.033	−2.859

表 15 において，$logit_1$ と $logit_2$ の差は一定ではない。

表 15　累積ロジスティックモデルの結果

	x_1	x_2	x_3	y	$logit_1$	$logit_2$	p_1	p_2	p_3	判別結果
1	62.4	51.7	51.8	3	-1.73110	0.52173	0.15045	0.47711	0.37245	2
2	48.3	47.9	43.8	1	-0.66717	1.78256	0.33913	0.51688	0.14399	2
3	61.1	37.9	48.6	1	-0.73442	1.16342	0.32423	0.43773	0.23805	2
4	48.4	42.3	37.4	2	-0.21090	2.50036	0.44747	0.47670	0.07583	2
5	48.0	57.1	61.1	3	-1.51999	-0.02643	0.17946	0.31393	0.50661	3
6	50.0	53.2	43.9	2	-1.09058	1.60495	0.25151	0.58120	0.16729	2
7	48.0	36.0	34.9	1	0.25144	2.88447	0.56253	0.38454	0.05293	1
8	51.8	53.0	68.9	3	-1.57096	-0.78090	0.17208	0.14205	0.68587	3
9	36.5	32.4	35.2	1	1.04064	3.28627	0.73897	0.22498	0.03605	1
10	33.1	47.5	43.8	1	0.10410	2.25494	0.52600	0.37907	0.09492	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
99	55.9	55.7	61.5	2	-1.82467	-0.27399	0.13888	0.29305	0.56807	3
100	68.7	64.8	59.0	2	-2.99317	-0.63074	0.04774	0.29961	0.65266	3

正判別率は 52% である。

表 16　累積ロジスティックモデル判別による判別結果

	1	2	3	合計
1	13	16	2	31
2	8	21	10	39
3	3	9	18	30